

Customer Churn Analysis and Prediction



Group -7

Members :

- Trupti Pendharkar
(IIT2018097)
- Puja Kumari
(IIT2018191)
- Prabha Kumari
(IIT2018195)

Under the supervision of :-

Dr. Pavan Chakraborty

Problem Statement

- Customer churning which can also be referred as customer attrition refers to the situation when customers in the bank tend to leave that bank.
- The cost of finding a new customer is much higher than that of maintaining old customers.
- To resolve the problem the computing age will help in the prediction of attrition.
- Data Mining and Machine Learning Techniques can be used on the data available by analyzing data from different viewpoints such as cause of attrition, its frequency , age of a customer, gender , rate of interest, geography etc. to create a model to predict if an customer is going to leave the brand or not for given user details.

Introduction

- In the era of big data, customer churn is a big problem faced by banks in the increasingly competitive market.
- As customer churning is rising with years and as it has a direct negative impact on the revenue of the bank. Customer churn may be because of several reasons like some other bank providing financial services at low charges or due to low interest rates or due to location of bank branch etc.
- So in order to stop this, bank need to prepare a prediction model in advance which would be able to predict the behaviour of the customer in advance about the customer that if he/she is going to leave the bank or not

In this paper we have used different models like ANN, XGBoost, Stacking Classifiers and presented the analysis of the different results obtained.

So our aim in this project is to create a model and then create a UI which will take the details of the employee as input and predict the output regarding attrition .

Motivation

- Nowadays, there are almost 1.5 million customers that are churning in a year that is rising every year.
- The Banking industry faces challenges to hold clients. The clients may shift over to different banks due to reasons like better financial services at lower charges, bank branch location, low-interest rates, etc.
- Thus, prediction models are utilized to predict clients who are probably going to churn in the future. Because serving long-term customers is less costly as compared to losing a client that leads to a loss in profit for the bank. Also, old customers create higher benefits and provide new referrals.

Literature Survey.....



Research Paper 1 : [\[1\]](#)

Paper title : Analysis and prediction of bank user churn based on ensemble learning algorithm

Name of Conference : 2021 IEEE International Conference on Power Electronics, Computer Applications (ICPECA), 22-24 Jan. 2021, Shenyang, China

Objective : The purpose of this paper is to analyze the quarterly user data of banks, and establish user churn prediction model by using ensemble learning so as to improve the accuracy of prediction, so as to achieve the purpose of helping banks save costs

Methodology : They used algorithm such as Catboost, Lightgbm, Random Forest to build their model.

Result : At the end of each quarter, the customer churn rate with a large amount of deposit or financial products is quite low, so they should focus on those users who have little deposit.



Research Paper 2 : [\[2\]](#)

Paper title : Application of Machine Learning and Statistics in Banking Customer Churn Prediction

Name of Conference : 2021 8th International Conference on Smart Computing and Communications (ICSCC), 06 September 2021, Kochi, Kerala, India

Objective : They are willing to make a website which is useful for the bank managers and decision makers of the bank to get an idea of those customers who are likely to leave the services of the bank in future and can retain them by formulating some new policies.

Methodology : They used Machine Learning Techniques like SVM and Statistical Analysis like Analysis of Numerical Data and Analysis of Categorical Data. Also they used 'Flask' framework to create the web application along with 'HTML' and 'CSS'.

Result : The model predicts the probability of the customer's leaving the bank and continuing the services of bank. The accuracy is around 84.15 %.

Research Paper 3 : [\[3\]](#)

Paper title : Customer Churn Analysis and Prediction in Banking Industry using Machine Learning

Name of Conference : 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC), 6-8 Nov. 2020, Waknaghat, India.

Objective : They are aimed to use different models of machine learning to the bank dataset to predict the probability of customer who is going to churn. The comparison in terms of performance like accuracy, recall, etc. is presented.

Methodology : They used algorithm such as Logistic regression (LR), decision tree (DT), K-nearest neighbor (KNN), random forest (RF).

Result : They observed that stratified and cross validation performs better in each case among all classifiers. But DT classifier has a .4429 recall value and 85.20% accuracy that is better as compared to others.



Research Paper 4 : [\[4\]](#)

Paper title : Machine Learning Based Customer Churn Prediction In Banking

Name of Conference : 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 5-7 Nov. 2020, Coimbatore, India

Objective : In this paper, a method to predicts the customer churn in a Bank, using machine learning techniques, is proposed.

Methodology : The KNN, SVM, Decision Tree, and Random Forest classifiers are used in this paper. Also, some feature selection methods have been done to find the more relevant features and to verify system performance.

Result : Result shows that the DT and RF classifiers accuracy increased after oversampling, but there is no change in KNN accuracy with regard to oversampling and SVM is not suitable for huge amounts of data. For KNN accuracy is 81.65%, SVM : 79.63%, DT-78.99%, RF-85.18% .



Research Paper 5 : [\[5\]](#)

Paper title : An Enhanced Bank Customers Churn Prediction Model Using A Hybrid Genetic Algorithm And K-Means Filter And Artificial Neural Network

Name of Conference : 2020 IEEE 2nd International Conference on Cyberspace (CYBER NIGERIA), 23-25 Feb. 2021, Abuja, Nigeria

Objective : They are proposed to use data mining techniques ANN and shown the performance of model .

Methodology : They used Artificial Neural Networks (ANNs) and two filters were applied to the data, the Genetic Algorithm (GA) and K-means filter .

Result : The results show that the training performance improved as the noise in the data reduces while the testing results were not improved with filtering.



Research Paper 6 : [\[6\]](#)

Paper title : Prediction of Customer Status in Corporate Banking Using Neural Networks

Name of Conference : 2020 International Joint Conference on Neural Networks (IJCNN)
28 September 2020, Glasgow, UK

Objective : This paper presents a computer system that is based on the application of artificial neural networks and support vector machine and used to predict the future status of corporate banking clients.

Methodology : They used two different classifiers and compared their result : a multilayer perceptron and a support vector machine.

Result : This study shows that the data mining techniques based on proper definition of input attributes and application of artificial neural networks provides a good tool for supporting the prediction of customer behaviour in corporate banking.

Dataset Description


[\[Dataset\]](#)

We will use the dataset available at kaggle named “ Bank Customer Churn data. Link ; [Dataset](#)

- Dataset consist of total 14 features
- Total record in dataset : 10,000

First ten features of the dataset are described as follows:

1. CustomerID : Describes Id of customer (Numerical Value)
2. Surname : Target variable
3. Credit Score : Integer Value.
4. Geography : Location Of Customer
5. Gender : String value

- 
-
6. Age : Numerical value
 7. Tenure : Numerical value
 8. Balance: Numerical Value
 9. HasCrCard: Binary Value
 10. IsActiveMember: Binary Value

And there are many more important features like estimated salary , exit status etc.



Language and tools

- Language: Python
- Libraries: numpy,pandas
- Web Interface using Flask



Methodology

After Data Collection we are going to perform 3 major steps . These are as follows

1. Exploratory data analysis.
2. Training Model.
3. Performance Analysis of various models.



1. Exploratory Data Analysis

1. Data exploration is the core part of a DM and ML project.
2. In EDA we have found that there are no missing values also we have dropped the irrelevant features from the dataframe like in our case roll no., customer id, and surname are irrelevant so we drop those features
3. Features like Geography and Gender were transformed into numerical variables.



Results we get in Data analysis were like:

1. Customers with 3-4 products have higher chances to churn.
2. Customers lying in the Age-Gap of 40-70 have higher chances to churn.
3. Customers with credit score less than 400 have higher chances to churn

After analysis we have done feature scaling using the MinMax scaling algorithm.

'Credit Score', 'Age', 'Tenure', 'Balance', 'Estimated Salary' were some of the features which were scaled down.



SMOTE(Synthetic Minority Oversampling Technique):

As our training data was highly imbalanced therefore we have performed Oversampling on Minority Classifier:

This algorithm works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample at a point along that line.

1. a random example from the minority class is first chosen.
2. Then k of the nearest neighbors for that example are found (typically $k=5$).
3. A randomly selected neighbor is chosen and a synthetic example is created at a randomly selected point between the two examples in feature space.



2. Algorithms for training:

We have observed that most of the researchers who have performed attrition analysis and prediction have either used Decision trees, or Random Forest, or Logistic Regression or Naive bayes.

We know that when it comes to efficient classifiers ANN and xgboost have always proved themselves most of the times in terms of accuracy.

So the models which we have tried in the project are ANN,XgBoost and Stacking Classifier for the later one we have used ANN and Xgboost as sub models and Xgboost as meta model.



1. XgBoost:

One of the best known ensemble technique and shows great performance and speed among all tree based ML algorithms.

It uses L1 and L2 regularization to prevent overfitting.

It has built in cross validation function

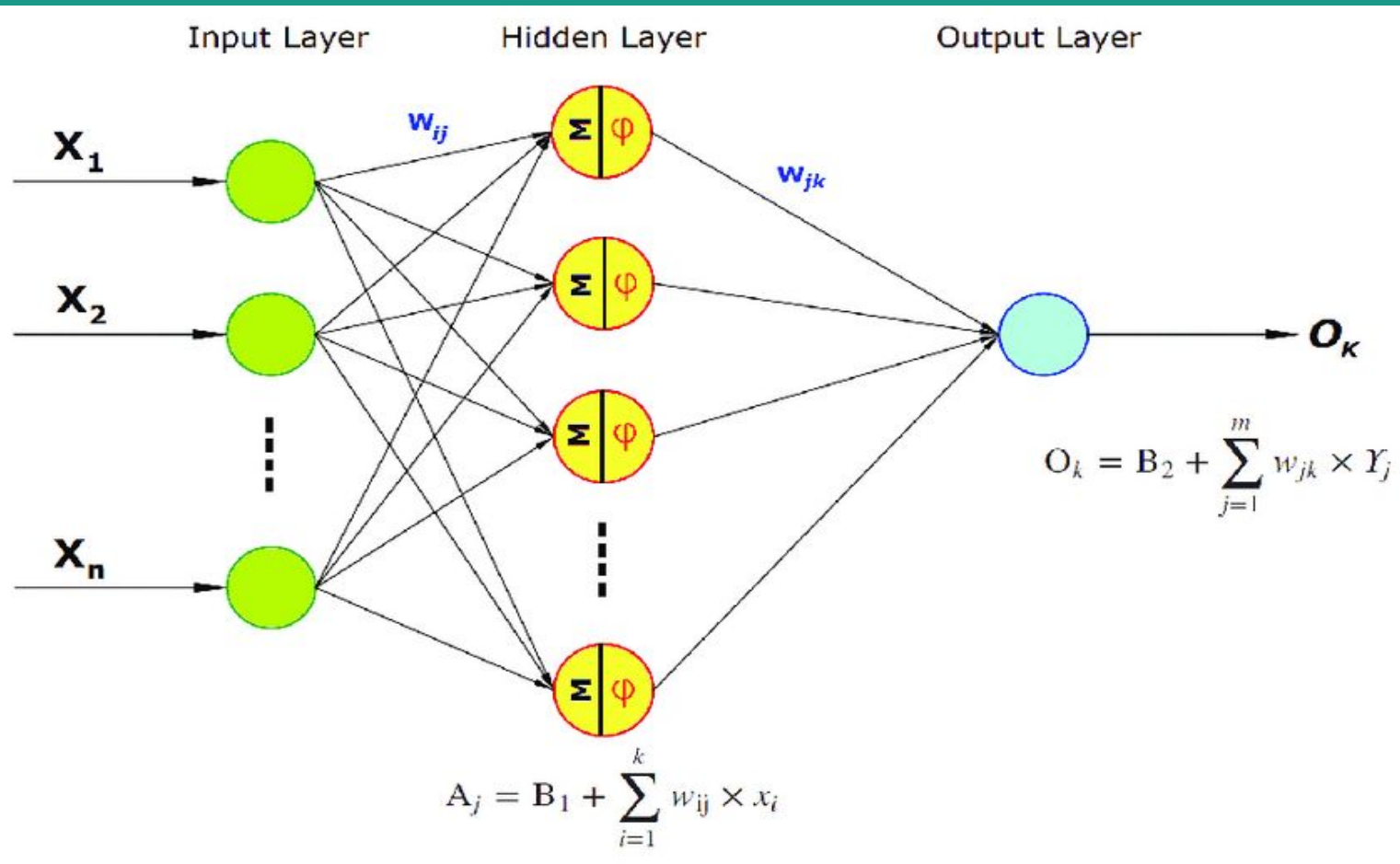




2. ANN:

We have also used feed Forward ANN also known as MLP(Multilayer perceptron) as one of our model . It works on the concept of weight optimization using back propagation algorithm.

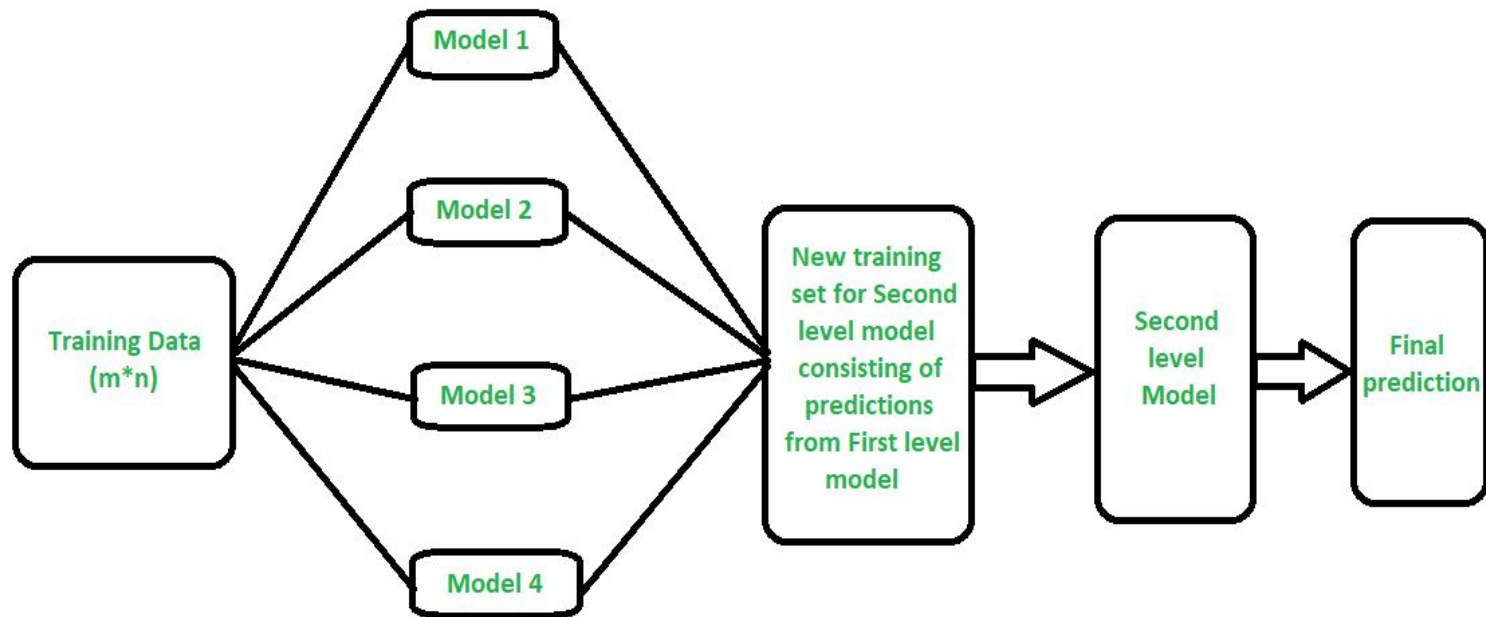
It consists of input ,hidden and output layers and each layer is fully connected to next layer.





3. Stacking Classifier

The simplest form of stacking can be described as an ensemble learning technique where the predictions of multiple classifiers (referred as level-one classifiers) are used as new features to train a meta-classifier. The meta-classifier can be any classifier of our choice.



Results (Without SMOTE)

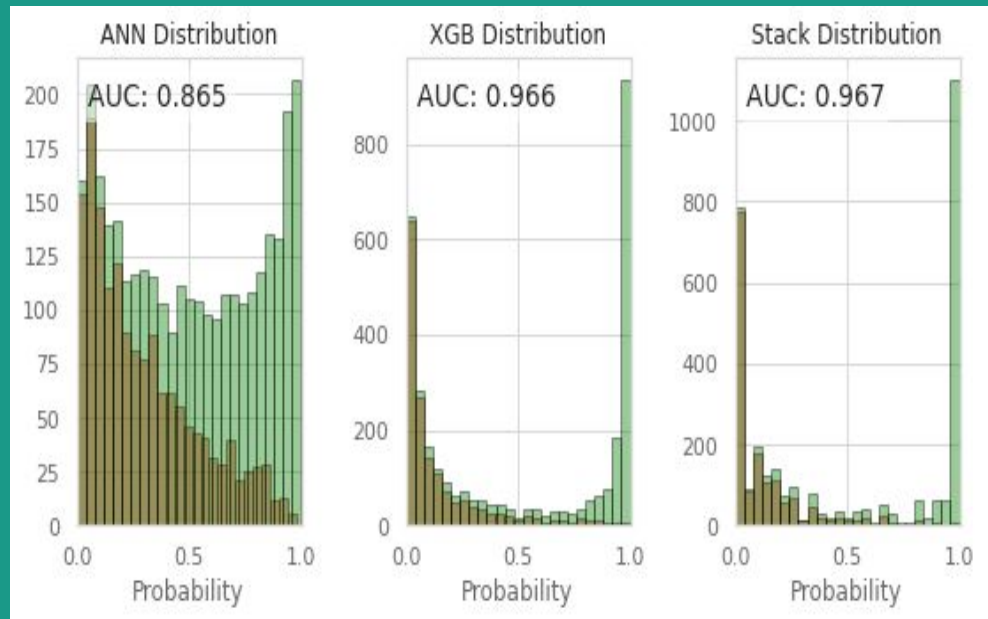
Metrics	ANN	XGBoost	Stacking Classifier
Accuracy(auc score)	0.865	0.966	0.967
Precision	0.789	0.919	0.930
Recall	0.775	0.881	0.874
F1 score	0.782	0.900	0.901

Results (When SMOTE was Applied)

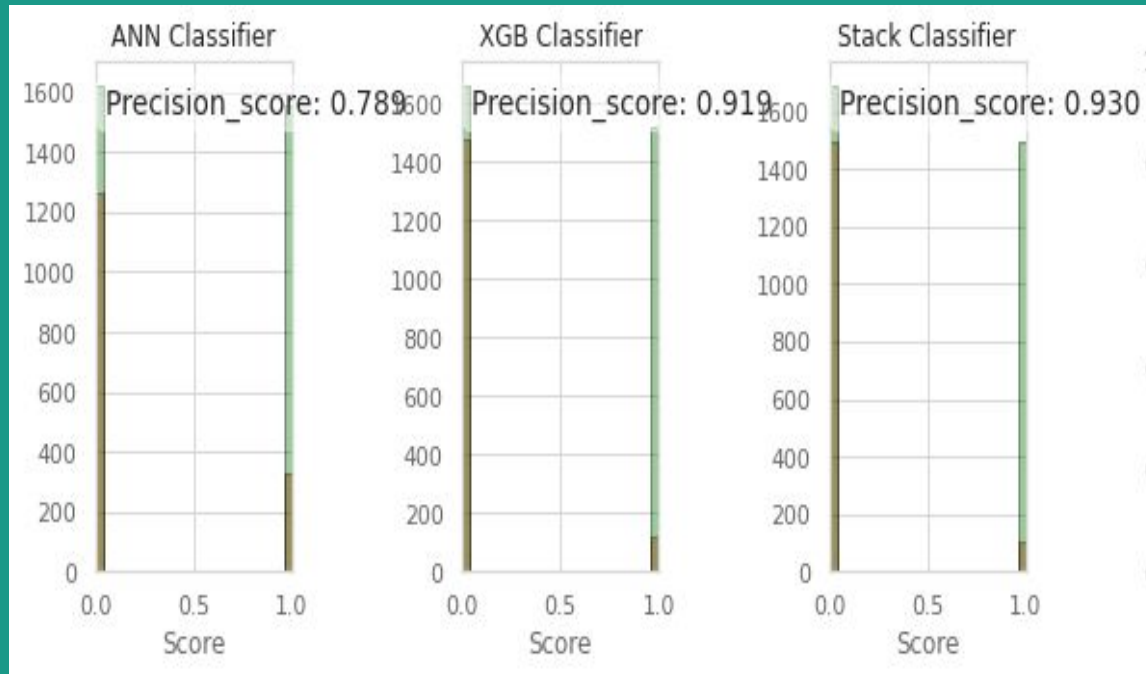
We have analysed the result on the basis of different metrics on our final data

Metrics	ANN	XGBoost	Stacking Classifier
Accuracy(auc score)	0.865	0.966	0.967
Precision	0.789	0.919	0.930
Recall	0.775	0.881	0.874
F1 score	0.782	0.900	0.901

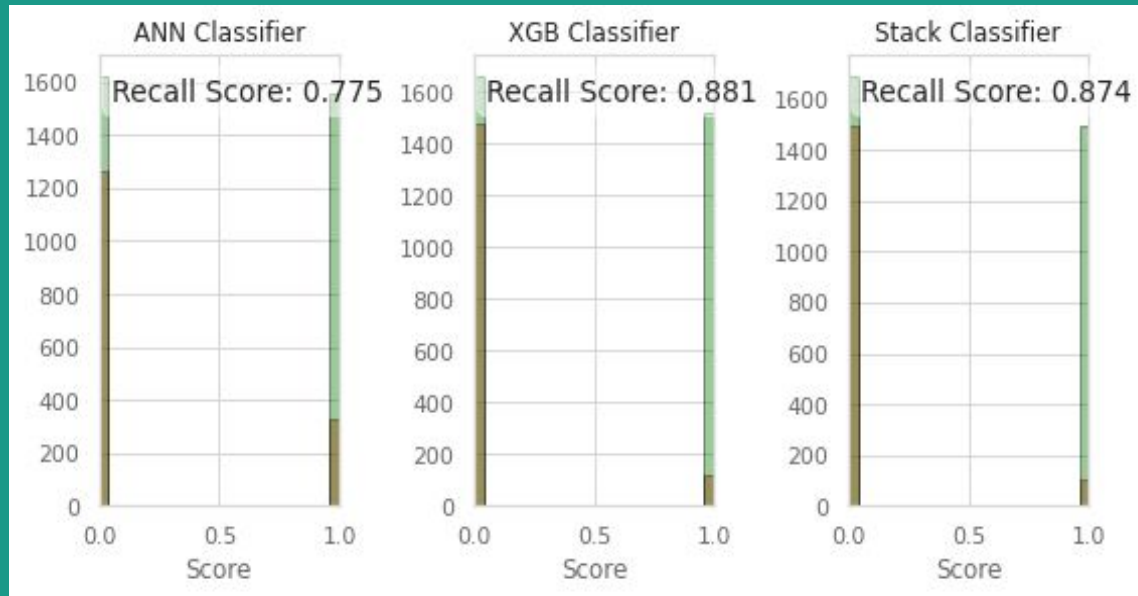
Accuracy score



Precision score



Recall score



F1 score





Design of Web Application

After analyzing the results on various metrics Stacking Classifier proved better on majority of them . So for our web application we have used Stacking classifier as our model.

For building the web application we have used HTML, CSS for frontend, and Flask framework for backend and deployed our application on heroku.

Link: <https://mini3-churnpred.herokuapp.com/>

Front page of application:

mini3-churnpred.herokuapp.com/predict

Banking Churn Prediction

700
France
Female
39
1
0
2
0
0
93826

Submit

Customer Churn rate: 9.67%

Activity Schedule:



Steps	Time Required	Predicted Date and time
Requirement Verify	Done	5th September 2021
Project Planning	Done	8th September 2021
System Design	Done	20th September 2021
Details Design	Done	30th September 2021
Coding	Done	15th October 2021
Debugging and coding	Done	30th October 2021
Testing	Done	10th November 2021
Documentation and Final	Done	20th November 2021



Conclusion

By Applying SMOTE techniques on our data results are improved and also we have used stacking classifier as our final model for the application since it has performed better on the majority of the metrics discussed above so our application which is deployed on the heroku uses stacking classifier model.



References

[1]. Saadat M Alhashmi, "Towards Understanding Employee Attrition using a Decision Tree Approach", 2019 International Conference on Digitization (ICD)

[Towards Understanding Employee Attrition using a Decision Tree Approach](#)

[2]. Richard Joseph, Shreyas Udupa, Sanket Jangale, Kunal Kotkar, Parthesh Pawar, "Employee Attrition Using Machine Learning And Depression Analysis", 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)

[Employee Attrition Using Machine Learning And Depression Analysis](#)

[3]. Rachna Jain, Anand Nayyar, "Predicting Employee Attrition using XGBoost Machine Learning Approach", 2018 International Conference on System Modeling & Advancement in Research Trends (SMART)

[Predicting Employee Attrition using XGBoost Machine Learning Approach](#)



[4]. Sepideh Hassankhani Dolatabadi, Farshid Keynia, “Designing of customer and employee churn prediction model based on data mining method and neural predictor”,2017 2nd International Conference on Computer and Communication Systems (ICCCS)

[Designing of customer and employee churn prediction model based on data mining method and neural predictor](#)

[5]. A Rohit Hebbar, Sanath H Patil, S. B Rajeshwari, S S M Saquaf, “Comparison of Machine Learning Techniques to Predict the Attrition Rate of the Employees”,2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)

[Comparison of Machine Learning Techniques to Predict the Attrition Rate of the Employees](#)

[6]. Sandeep Yadav, Aman Jain, Deepti Singh, “Early Prediction of Employee Attrition using Data Mining Techniques”, 2018 IEEE 8th International Advance Computing Conference (IACC)

[Early Prediction of Employee Attrition using Data Mining Techniques](#)

Thank you !!